The Effects of Auditory, Visual, and Gestural Information on the Perception of Mandarin Tones

by

Beverly Joyce Ching-Sum Hannah

BA (Honours), Simon Fraser University, 2012

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Arts

in the
Department of Linguistics
Faculty of Arts and Social Sciences

SIMON FRASER UNIVERSITY
Summer 2017

Approval

Name:	**Beverly Hannah**
Degree:	**Master of Arts [Linguistics]**
Title:	**The Effects of Auditory, Visual, and Gestural Information on the Perception of Mandarin Tones**
Examining Committee:	**Chair: Chung-hye Han** Professor
	Yue Wang Senior Supervisor Associate Professor
	Ashley Farris-Trimble Supervisor Assistant Professor
	Allard Jongman External Examiner Professor Department of Linguistics University of Kansas
Date Defended/Approved:	August 4, 2017

Ethics Statement

The author, whose name appears on the title page of this work, has obtained, for the research described in this work, either:

a. human research ethics approval from the Simon Fraser University Office of Research Ethics

or

b. advance approval of the animal care protocol from the University Animal Care Committee of Simon Fraser University

or has conducted the research

c. as a co-investigator, collaborator, or research assistant in a research project approved in advance.

A copy of the approval letter has been filed with the Theses Office of the University Library at the time of submission of this thesis or project.

The original application for approval and letter of approval are filed with the relevant offices. Inquiries may be directed to those authorities.

Simon Fraser University Library
Burnaby, British Columbia, Canada

Update Spring 2016

Abstract

In multimodal speech perception, strategic connections between auditory and visual-spatial events can aid in the disambiguation of speech sounds. This study examines how co-speech hand gestures mimicking pitch contours in space affect non-native Mandarin tone perception. Native English as well as Mandarin perceivers identified tones with either congruent (C) or incongruent (I) Audio+Face (AF) and Audio+Face+Gesture (AFG) input. Mandarin perceivers performed at ceiling rates in the Congruent conditions, but showed a partially gesture-based response in AFG-I, revealing that gestures were perceived as valid cues for tone. The English group's performance was better in congruent than incongruent AF and AFG conditions. Their identification rates were also highly skewed towards the visual tone when gesture was presented in the AFG compared to AF conditions. These results indicate positive effects of facial and especially gestural input on non-native tone perception, suggesting that crossmodal resources can be recruited to aid auditory perception when phonetic demands are high.

Keywords: lexical tone; pitch; Mandarin Chinese; crossmodal; gesture; audiovisual

Dedication

To my dearest Eleanor and Apollo,

Hello to the future!

If you're reading this, you're probably all grown up by now, and that is good.

You've been asking me about time capsules lately. So here's one for you to find one day. This one won't go in the ground, but hopefully, it will be kept for a very long time.

I look at you both, and I can't believe that you're real. You're here. You are people. Actual tiny people, with *very* distinct personalities. Your feet keep growing at alarming rates. You've just become each other's best playmates and worst competition, depending on the minute. Apollo, you're obsessed with football, cars, buses, and vacuums. You've just hit that stage where you've started to say new words every day. "Football" was one of your earliest words, go figure. You say thank you for everything and to everything, even to the crosswalk light when it changes, and it's adorable. Eleanor, you've just learned to swim and ride a bike and pick salmonberries and tell knock knock jokes. You just said thank you in public for the second time EVER, after the three of you got haircuts (Apollo's first!) yesterday. We were so proud of you. This morning, we brought peanuts to Stanley Park to feed the crows and geese and seagulls, and then we looked for crabs and shells at the beach. Then while I wrote, Daddy took you exploring down the ravine behind Grandma's with the inflatable dinghy and you played pirates and "paddled" around the puddle-deep waters and dug for treasure on the sandy shores. We listen to Peppa Pig in the car every day on the way to and from daycare at SFU. It's my favourite show. They're hilarious.

Raising the two of you has been an expansion of my life in depth and colour and texture and richness in all directions and dimensions beyond every expectation and limit I imagined possible. Complete sensory processing overload. I've felt so overwhelmed and broken these past four years. To me, this thesis is merely endpaper to the story of finding ourselves as a family. You two are my most treasured stories. I fear that I won't be able to remember very much of this later, and I'm guessing you won't either. But I do hope we will try to remember the good bits. Let's reminisce about the good things often, and maybe they'll stick. I hope that in the future, there will be places and smells and songs and foods that will transport you back to these magical early days. They really are magical. Already, when I close my eyes and stay very still, I can hear bubbles in my heart that weren't there before. Sparkling fizzing peals of giggles. Beautiful, mischievous, tickly giggles, bubbling over, flooding my heart with fountains of sudden shiny rainbows,

pop! pop! pop! pop! pop!

…and I cannot help but be filled with the warmth of a thousand sunrises.

I have no idea why we are here, but I feel so lucky that we get to be here together.

To my dearest Steve,
I can't believe we get to fall apart together.

All my love,
xoxo Mummy

Acknowledgements

For her patience, kindness, understanding, and guidance, I thank my supervisor, Dr. Yue Wang. It has been an incredible opportunity to work with you at the LABlab for the past six years through my undergrad, 2 kids, 3 conferences, 4 posters, 5 publications, 6 lab projects, and where were we? Oh right, an MA. You rock.

For the sparkles they leave in their wake, I thank my compatriots at the Language and Brain Lab at Simon Fraser University, past and present. You're all shooting stars, the best and the brightest, and I love celebrating with you as you achieve your dreams.

For their assistance on this project, I would specifically like to thank Anthony Chor, Courtney Lawrence, Daniel Chang, Danielle Weber, Eleanor Hendriks, Elysia Saundry, Jennifer Williams, Katelyn Eng, Katy Thompson, Keith King-Wui Leung, Sylvia Cho, and Xiaojie Zhao from the Language and Brain Lab.

I also would like to extend my sincerest appreciation to Jiguo Cao and Yunlong Nie from the Data Science Lab at Simon Fraser University for their assistance with data analysis, and for the tangential but thoroughly entertaining discussions about interesting problems in statistics!

To Dr. Ashley Farris-Trimble, thank you for serving on my supervisory committee, and for not calling on me too much when I couldn't stay awake during your Phonology class. The grad school student by day, swashbuckler of tiny fire-breathing dragon by night thing, was, as they say, an adjustment. 💣

Mahalo nui loa to my academic grandparents Drs. Allard Jongman and Joan Sereno for your valuable feedback on my manuscript, for serving on my examining committee, and for sharing your aloha spirit. Until we meet again!

This study was supported by a Social Sciences and Humanities Research Council of Canada (SSHRC) research grant to Dr. Yue Wang entitled *Multi-lingual and multi-modal speech perception, processing, and learning* (SSHRC Insight Grant 435-2012-1641).

Portions of this study were presented at the 5th Joint Meeting of the Acoustical Society of America and Acoustical Society of Japan, Honolulu, HI, November, 2016.

Table of Contents

List of Tables

List of Figures

List of Acronyms

AV	Audio-Visual
AF	Audio-Facial
AFG	Audio-FacialGestural
C	Congruent
I	Incongruent
L1	First Language or Native Language
L2	Second Language

Glossary

Pitch	Human perception of acoustic frequency (Hz). For this experiment, pitch refers to the relative range of the voice, using descriptors such as low, mid, and high
Pitch Contour	The spatial description of pitch as it changes over time. This experiment describes the pitch contours of lexical tones with terms such as *high falling tone* and *mid-rising tone*
Lexical Tone	Word-level pitch or pitch contour that distinguishes between words containing the same phonemes
Gesture	Communicative movements of the hand and arm that convey independent, redundant, or complementary information to that contained in speech.
Audiospatial	The concept that sound is perceived, processed, and represented both acoustically and spatially in the mind
Modality	For this study, Modality refers to the type of stimuli perceived as input during the experiment – either with facial or with facial and gestural information
Multimodal	The presence of more than one sensory input or channel of information during perception
Crossmodal	The connection of two or more types of sensory inputs determined to be from the same event
Congruency	For this study, Congruency refers to whether the auditory tone information and visual tone information in the gesture and/or face match in stimulus input
Visual Saliency	For this experiment, visual saliency refers to the extent to which visual components, such as facial or gestural movements, stand out from the background
Cue Weighting	The extent to which perceivers use information from one sensory input over another, especially when experiencing a perceptual conflict
Speaker	In this experiment, a participant who produces speech tokens
Perceiver	In this experiment, a participant who listens to and/or watches speech and gesture tokens
Native	Processing speech sound and syntactic categories consistent with a population who were exposed to Language X since birth.

Non-native	A non-native speaker of language X would possess native speech sounds and syntactic categories from their own Language Y.

Chapter 1. **Introduction**

From infancy onward, perceivers of multisensory stimuli are continually tasked with solving the crossmodal binding problem (Spence, 2011), where sensory signals from the same event must be matched for processing. This requires perceivers to determine that the relevant inputs are semantically congruent, spatiotemporally correspondent, and crossmodally associated (Ernst & Bülthoff, 2004; Fujisaki & Nishida, 2007; Spence, 2011). Before the signals can be bound, however, the perceiver must first establish that the signals have indeed arisen from the same event. In the context of everyday language use, crossmodal binding is not usually problematic unless the auditory speech stream is rendered unintelligible by background noise or other barriers to real-time language decoding and comprehension. When irrelevant acoustic information cannot be easily filtered out during processing, crossmodal mapping of facial movements to acoustic features can support the processing of speech events based on their visible articulatory configuration and timing cues. More peripherally, co-speech manual gestures can also support speech processing by marking beats in time, pointing the way, illustrating content, or by metaphorically translating sounds into spatial adjectives such as *high* and *low* (McNeill, 1992; Marks, 1987). When these spatially described sounds move and change through time as *pitch contours,* they can be represented dynamically as in the *level*, *rising*, *dipping*, and *falling pitch gestures* of the present study. With these additional cues recruited, crossmodal binding can take place. After the relevant sensory signals have been bound to the same event, the integration of auditory and visual streams can occur, resulting in either perceptual fusion or crossmodal conflict. The likelihood of fused versus conflicting percepts can vary greatly depending on perceiver experience (Ernst, 2007; Parise & Spence, 2009). The perceivers in this study differ in linguistic experience: naïve versus native in their knowledge of the test language used in the experiment. Thus, it will be the reliability of each input, as judged by the perceivers, that will reveal differing patterns of perceptual fusion or conflict.

The present study tests the strength of crossmodal association between acoustic pitch information and visuospatial tone information. We explore whether gesture can bias pitch perception in a linguistically significant context, and more broadly, how visual cues may be differentially utilized by native versus non-native perceivers. Experimental stimuli

consist of videos of speakers producing Mandarin lexical tones with and without tone gestures, aurally masked by background noise. Participants are then tasked with identifying the tones they perceived. In a manner akin to the early studies of crossmodal auditory-visual correspondences (Bernstein & Edelstein, 1971; Marks, 1987) and the more recent studies of gesture and speech integration (Kelly, Ozyürek, & Maris, 2010), the experimental design manipulates the semantic congruency of the tonal information and presents audiovisually tone-congruent and tone-incongruent stimuli to native (Mandarin) and non-native (native English) perceivers. In non-native perceivers, we predict that the tone identification task may induce strong cross-modal linking, resulting in more visually-influenced responses despite the unreliability of the visual information. Native Mandarin perceivers, however, would be expected to show less of a visual effect, relying less on the visual information, as they already have firmly established tone categories as part of their native phonetic inventories.

This research may be of interest to the language teaching and learning community at large for its insight into how tonal perception can be altered with concurrent exposure to different types of visual information, be it through facial or gestural cues. Empowered with this information, they will be able to manipulate these perceptual influences to their advantage as part of their teaching and learning strategies.

1.1. Facial Cues for Speech

Previous research has observed auditory-face (AF) vowel binding in infants as young as two months of age (Kuhl & Meltzoff, 1982, 1984; Patterson & Werker, 2003). During critical times for learning their native language, infants have been found to shift their gaze patterns from looking primarily at the speaker's eyes to the speaker's mouth (Lewkowicz & Hansen-Tift, 2012). In contrast, adults have been found to generally look at the speaker's eyes, only looking more at the speaker's mouth when the speech signal is masked by increasingly louder noise (Vatikiotis-Bateson, Eigsti, Yano, & Munhall, 1998). Similarly, complementary visual information can be recruited to improve signal quality when comprehension conditions are less than ideal, such as when learning a second language, listening to non-native speakers, or perceiving speech in noise. Visual cues gained from watching a speaker's face have been shown to enhance the processing of segmental speech in native (L1) and non-native (L2) perception (Summerfield, 1983; Jongman, Wang, & Kim, 2003; Davis & Kim, 2004; Wang, Behne, & Jiang, 2008; 2009);

however, the facial area that may provide the most visual benefit may depend on the type of information sought, such as the eyebrows and upper part of the face for prosody or the lower part of the face for word level information (Cavé et al., 1996; Lansing & McConkie, 1999; Swerts & Krahmer, 2008). The visual saliency of a speech sound must also be taken into account when evaluating the potential visual benefit over auditory-only perception. The most visually salient sounds offer the greatest opportunities for articulatory movements to contribute to speech perception (Hazan et al., 2006; Hazan, Kim, & Chen, 2010). Additionally, when conversing with non-natives or when communication is otherwise impaired, such as in the presence of background noise (Sumby & Pollack, 1954; Macleod & Summerfield, 1990; Nielsen, 2004) or when repeatedly asked for clarification, speakers may modify their speaking style to produce clear speech with exaggerated visual cues for jaw displacement, lip stretching, lip rounding, and duration, increasing the visual saliency of their speech sounds in comparison to plain speech, ostensibly for the perceiver's benefit (Tang et al., 2015). The McGurk effect (McGurk & Macdonald, 1976) is one well known instance where high visual saliency can result in perceptual fusion for Audio-Facial (AF) stimuli with mismatched place of articulation cues. Increased perceptual fusion and, by extension, visual reliance, has been shown to occur in perception of non-native McGurk stimuli (Sekiyama & Tohkura, 1993; Chen & Hazan, 2007). However, the perceptual fusion of the McGurk effect has been shown to be negatively affected when task and attentional demands exceed perceptual load capacities (Alsius, Navarra, Campbell, & Soto-Faraco, 2005; Lavie 1995)

1.2. **Mapping Auditory Pitch to the Visual Domain**

In the general cognitive domain, the crossmodal association between auditory pitch and visual elevation, as well as between auditory pitch and spatial movement have been well established (Casasanto, Phillips, & Boroditsky, 2003). One linguistic example of crossmodal auditory and visual pitch association is in Mandarin Chinese, where lexical tone is sufficient to differentiate two otherwise identical syllables as a minimal pair. The level, rising, dipping, and falling tones are represented in the Pinyin Romanization system by tone shape diacritics that represent each tone's pitch height and contour (Xu, 1997; Chao, 1968; Lin, 1985). These Pinyin diacritics are introduced concurrently with tones in beginners' Mandarin classes for non-native learners, and are taught to native Mandarin speaking schoolchildren as part of their elementary studies as well. The literature abounds

with suggestions for helping learners to form crossmodal associations between auditory pitch and visual height for tone in this way, such as by writing pinyin on a musical staff in the shape of the tonal contours (Lin, 1985), or by learning Pinyin words with visual contour diagrams in lieu of the more arbitrary 1-4 numerical association system (Liu et al., 2011).

1.3. **Facial Cues for Tone**

There is consensus among previous studies that visual cues for tone are available in the facial region of the speaker (Attina, Gibert, Vatikiotis-Bateson, & Burnham, 2010; Burnham, Ciocca, & Stokes, 2001; Burnham, Lau, Tam, & Schoknecht, 2001; Burnham, Reynolds, Vatikiotis-Bateson, Yehia, & Ciocca, 2006; Chen & Massaro, 2008; Mixdorff, Hu, & Burnham, 2005; Mixdorff, Wang, & Hu, 2008; Smith & Burnham, 2012); however, the salience of these cues may depend on whether words are produced in citation or in phrasal contexts. In citation form, spatiotemporal cues for duration have been shown to be one of the more reliable indicators of tone, with falling being the shortest, and dipping measuring the longest. This durational difference between tones does not hold in sentential contexts, so in these cases, visual cues for syllable duration cannot be relied upon to help the perceiver distinguish between tones (Attina et al., 2010). Aside from duration, the head and neck movements that physiologically modulate pitch also produce visible cues for tone (Attina et al., 2010; Chen & Massaro, 2008; Smith & Burnham, 2012). Interestingly, these cues have been observed to be universally available, and more readily utilized by non-native speakers than native speakers (Smith & Burnham, 2012). Indeed, such cues may not be utilized by native speakers until they are brought to their attention (Chen & Massaro, 2008). However, when perceptual and cognitive load are increased, such as when identifying both tonal and segmental components of a word embedded in auditory noise, visual information fails to aid tone identification to the same extent as it does for segmental information (Mixdorff et al., 2005, 2008).

Since facial information has been shown to be helpful to perceivers by providing complementary visual cues when phonetic demands are high, if perceivers are forced to shift their attention to the visual modality for lexical tone perception, the presence of facial information for tone should be reflected in their tone identification patterns, as long as participants do not experience cognitive overload while completing their tasks. The uniqueness of the present experiment lies in the type of manipulation used to induce visual reliance during tone perception, namely, an Auditory-Facial incongruent modality of

presentation combined with embedding the acoustic signal in noise. In this modality, segmental information remains crossmodally congruent but tonal information crossmodally incongruent. Combining these stimuli with a single output task of tone identification allows participants to focus on tone, while being affected by visual cues for tone.

1.4. **Perceiving Gesture**

Manual gestures can enhance speech perception by providing focus cues and visual representations of the speaker's message, alternately relieving the speech modality of some of its semiotic burden of communication, or emphasizing the content of speech by providing redundant cues (Hostetter, 2011). It is this tightly bound interaction of speech and gesture that has led to the proposal of the *integrated systems hypothesis* (Kelly et al., 2010), which not only posits an influence of gesture on the perception of speech sounds and vice versa, but that such an interaction is mandatory, in that neither modality can be processed without considering the other, when both are available to the perceiver.

Previous studies have indicated that perceiving gestures with speech can indeed have significant effects on the perception of speech. For example, in L1 perception, beat gestures have been shown to alter the perception of word prominence (Krahmer & Swerts, 2007), and also to shift early sensory ERPs when integrated with speech, possibly providing additional parsing and focus cues for the perceiver (Biau & Soto-Faraco, 2013). For iconic gestures, when gestural content is manipulated to be incongruent with speech content, there is evidence for differential processing of these types of stimuli compared to when content is semantically congruent (Kelly, Kravitz, & Hopkins, 2004; Kelly et al., 2010; Kelly, Ward, Creigh, & Bartolotti, 2007; Kelly, Healey, Özyürek, & Holler, 2015). The present study will also be exploring the effects of semantic congruency on perception, but for a different type of stimuli – a crossmodal metaphoric representation of auditory pitch height and contour, or *tone gesture* (Eng, Hannah, Leung, & Wang, 2014; Morett & Chang, 2015).

1.5. **Gesture in L2**

Research into gesture and second language teaching has demonstrated that the simplified “teacher talk” register used by language instructors to engage students in classroom settings contains a substantially increased proportion of both representational and rhythmic gestures (Barnett, 1983; Gullberg, 2006; Lazaraton, 2004). In experimental studies however, gestures have not necessarily been found to be beneficial to the learner for acquiring non-native speech sounds. Neither observing nor producing beat gestures has been found to be particularly helpful for learning difficult phonemic contrasts of Japanese (Hirata & Kelly, 2010; Hirata, Kelly, Huang, & Manansala, 2014; Kelly, Hirata, Manansala, & Huang, 2014), while iconic gestures have only been found to aid word learning when word pairs containing the target phonemic contrasts were highly dissimilar in their surrounding segmental contexts, or when words were learned in isolation (Kelly & Lee, 2012; Kelly, McDevitt, & Esch, 2009). While beat gestures have not been shown to be effective at the word level, there is evidence for their effectiveness at the prosodic level, such as when used by L2 learners of English for practicing rhythm and syllabification in conversational settings (McCafferty, 2006), and also when used by instructors to demonstrate prosody at the discourse level, where the need to generate complex sentences often results in decreased intelligibility in L2 speech (Gluhareva & Prieto, 2016). At the discourse level, when the speaker is not constrained to a single type of gesture for experimental purposes, gestures have been shown to facilitate lexical access, improving comprehension over auditory-face and audio-only conditions, especially when comprehension is impaired by low language proficiency (Sueyoshi & Hardison, 2005). Together, these findings suggest that gestures are more effective for enhancing perception of non-native speech when difficult phonemic distinctions are not the primary focus of the exercise. When faced with highly demanding phonetic tasks, such as discriminating between phonemic durational cues in minimal pairs, additional gesture in all forms may prove to be too distracting while trying to make a fine-grained acoustic perceptual judgement.

1.6. **Pitch and Gesture**

For the present study, we are interested in seeing how perception of pitch can be influenced by manual gestures representing tone contour, where pitch and time are

mapped to vertical and horizontal movement. Capturing pitch in such gestures owes its inspiration to the illustrative aids of musical expression. Indeed, to create strong audiospatial connections, the Kodály music education system encourages early stage learners to kinesthetically engage in their experience of music using gestures and physical movements (Houlahan & Tacka, 2008). Music teachers are taught to enhance pitch perception using hand levels and diagrams of melodic contours (Apfelstat, 1988; Welch, 1985), and young singers are trained to improve their pitch accuracy using gestures (Liao and Davidson, 2007; Liao, 2008).

Experimentally, when participants have been given the freedom to represent stimulus sounds gesturally in a three-dimensional space, higher pitch has been generally found to correlate with higher elevation in space. It is interesting to note that in this free-form gesture production task, musicians have been found to be more consistent in their gesturing than non-musicians, which may provide insight into how the strength of crossmodal associations can develop with experience (Küssner, Tidhar, Prior, & Leech-Wilkinson, 2014). Furthermore, it has been claimed that pitch is audiospatial in representation, implying that pitch perception is inescapably affected in the presence of spatial information, as opposed to being a unimodal auditory percept. Experimentally, this has been shown to be the case, in that upward and downward gestures have been shown to bias pitch perception in the direction of the gesture. Moreover, when cognitive loads are increased with memory tasks, the pitch perception bias remains upon introduction of a verbal task, but disappears for a spatial task, indicating that pitch perception may be a lower priority usage for spatial resources when alternate, highly reliable modalities – such as acoustic pitch information – are available (Connell, Cai, & Holler, 2013).

1.7. **Pitch, Tone, Face, Gesture – Together in L2 Perception**

How might the findings from the aforementioned studies affect the perception of pitch in a linguistic context? As Mandarin tones are examples of linguistically significant word-level pitch, we once again return to pitch and its usage in speech communication. Mandarin tones in citation form are particularly well suited to gestural representation as they spatiotemporally and semantically bind auditory pitch, visual elevation, and duration into speech-gesture events, and are represented orthographically as diacritics in Mandarin Pinyin. Full body gestures such as reaching for the ceiling, slumping the shoulders, eyebrow raising, head nodding, and foot stomping can help to magnify pitch and contour

differences and aid learners in exploring the pitch range required for a tonal language (Chen, 1974; Tsai, 2011; Zhang, 2006). For the current study, we can quite straightforwardly translate these shapes into manual gestures according to the tone shapes of Chao (1968). Concerning the question of ecological validity of these tone gestures, a quick YouTube search of the term “Mandarin tones” presents an ever-evolving stream of enthusiastic, media-savvy teachers and learners demonstrating these tone gestures and sharing their experiences and opinions on what is helpful for learning to distinguish between tones. In the literature, two previous studies (Eng et al., 2014; Morett & Chang, 2015) have utilized these tone gestures to examine tone and word learning in Mandarin.

Eng et al. (2014) adapted the auditory tone training paradigm of Wang, Spence, and Jongman (1999) in order to train non-native perceivers to learn Mandarin tones using tone gestures. In this case, both gesture and no-gesture groups improved during and after training. However, slower learning trajectories were observed for more phonetically challenging sounds when paired with gesture than without. The results showing that additional visual resources in the form of tone gesture were more distracting than assistive to the learners is in line with the research on word-level iconic and beat gestures in L2 perception (Hirata & Kelly, 2010; Hirata et al., 2014; Kelly et al., 2014, 2009; Kelly & Lee, 2012), where gestures were found to aid learning of phonemic and word-meaning contrasts when the phonetic and perceptual demands were not too high.

A similar training study, Morett and Chang (2015), showed improved tone identification for words used during training while semantic gesture and no-gesture conditions did not. While promising, the pitch gesture group showed no additional benefit in a generalization post-test over the semantic and no-gesture training groups. However, the smaller set of four pitch gestures (one per tone) appeared to be more effective for short-term tone recall than the larger set of twelve semantic gestures (one per word). Similar results were found for word-meaning association learning, where the smaller pitch gesture set improved word-meaning learning, whereas the larger semantic gesture set did not. It may be the case that for word learning, the difference in gesture set size may have contributed to increased spatial memory load (Connell et al., 2013), suppressing the effects of the semantic gesture. It would be interesting to repeat their experiment with similar, size-matched sets of gestures in order to determine whether the relationships

reported between gesture and the learning of tone and word meanings can be attributed to other factors.

1.8. **Present Study**

The audiospatial representation of pitch described in Connell et al. (2013) remains unstudied in a linguistic context, leaving the following questions unanswered: can tone gestures serve as cross-modal links between auditory and spatial representations of tone, or are the tone gestures merely arbitrary mnemonic devices? Based on our ability to bind metaphoric spatial terms with auditory features such as low and high, the evidence for crossmodal association of auditory pitch and visual elevation and for audiospatial representation of pitch, the present study will be exploring similar questions regarding gesture's role in linguistic pitch perception. We propose investigating how crossmodal binding for pitch, perceived in a linguistic context as Mandarin lexical tone, occurs for native Mandarin speakers and non-native (native English) speakers when they are presented with Audio-Facial (AF) or Audio-FacialGestural (AFG) stimuli. The stimuli presented will be manipulated to be either tone-congruent (C), where the visual and auditory tone match, or incongruent (I), where the visual tone and auditory tone are mismatched. Perceivers will then identify the tones based on the input they perceived. For facial effects, we hypothesize that if perceivers are able to effectively incorporate facial tonal cues, they would more accurately identify tones with congruent (than incongruent) audio and facial input. For gestural effects, if perceivers are able to establish a cross-modal link between the acoustic and visuospatial pitch information, they would more accurately identify tones when gestural input is available, and would be more accurate with congruent (than incongruent) audio and gestural input. However, if such a link is arbitrary, we should instead find that congruent and incongruent input result in equal performance. In addition, within the AFG-incongruent (AFG-I) condition, visual weighting should increase when perceivers are presented with visually salient information. In the present experiment, the gestures in the AFG-I condition move across the screen concurrently with the speaker's facial movements and are thus more visually salient than the movements in the AF-I condition. For L1 and L2 effects, we expect that English perceivers would be more affected by facial and gestural input than Mandarin perceivers, as they need additional resources to process challenging L2 tones.

We undertake the current study with the assumption that the representation of pitch in tone is not arbitrary, and is crossmodally linked. Thus, when perceivers are presented with audiovisually tone-incongruent stimuli, the conflict of speech and gesture cues may produce unexpected patterns of tone perception, even for native speakers of Mandarin, who have firmly ingrained tonal categories as part of their native phonological inventories.

Chapter 2. **Methods**

This study was carried out with the approval of the Office of Research Ethics at Simon Fraser University with written informed consent from all participants.

2.1. **Perceivers**

Fifty-two native and non-native Mandarin perceivers participated in the perception experiment. The native perceiver group consisted of 26 native speakers of Mandarin (15 female) born and raised in northern China or Taiwan, aged 19-33 (mean: 24). The non-native perceiver group consisted of 26 native speakers of English (14 female) born and raised in Western Canada or the USA, aged 19-30 (mean: 24), with no prior tone language experience or formal musical training (Cooper & Wang, 2012). All perceivers reported normal hearing and normal or corrected to normal vision, and no history of speech or language disorders.

2.2. **Stimuli**

2.2.1. **Characteristics and Types of Stimuli**

Eight Mandarin words (2 syllables (*ye*, *you*) x 4 tones) were chosen as the experimental stimuli for this study. Eight additional words were used as tone familiarization stimuli (syllable *duo* x 4 tones) and task familiarization stimuli (syllable *ge* x 4 tones).

Two modalities were recorded for the target words: Audio + Facial, where the speaker's facial (mouth) movements were presented while speaking a corresponding target word; and Audio + FacialGestural, where the speaker made a matched tone contour shaped hand gesture while speaking a word. The stimuli were further edited to create two incongruent audio-visual stimulus types where the auditory tone input did not match the facial and gestural tone input (see 2.2.3 for details on stimulus editing). Thus, in total, four types of stimuli were developed, as illustrated in Figure 1: (1) congruent Audio and Facial tone input (AF-C), (2) incongruent Audio and Facial tone input (AF-I), (3) congruent Audio and FacialGestural tone input (AFG-C), and (4) incongruent Audio and FacialGestural tone input (AFG-I).

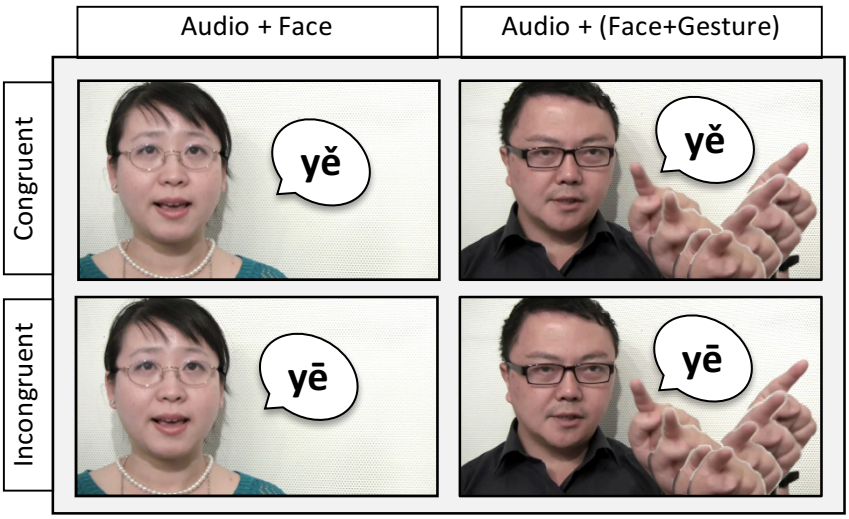

Figure 1 Four types of experimental stimuli, exemplified using syllable *ye* with Tone 1 (*yē*) and Tone 3 (*yě*)

(1) upper-left: congruent Audio and Facial tone input (AF-C): *yě*
(2) lower-left: incongruent Audio and Facial tone input (AF-I): audio *yē* + video *yě*,
(3) upper-right: congruent Audio and FacialGestural tone input (AFG-C): *yě*
(4) incongruent Audio and FacialGestural tone input (AFG-I): audio *yē* + video *yě.*

2.2.2. **Speakers and Recording**

The experimental stimuli were produced by two native Mandarin-speaking instructors (1 female aged 34, and 1 male who appeared of a similar age but declined to disclose his age) with experience teaching college-level introductory Mandarin classes. Two additional native Mandarin speakers (1 male aged 22, and 1 female aged 24) produced the audio-only tone familiarization stimuli. The speakers reported no history of speech or language disorders.

Audio-visual recordings were made in a sound-attenuated booth at the Language and Brain Lab at Simon Fraser University. For the AFG condition, the speakers were asked to simultaneously say each word in citation form while tracing a matching tone contour in the space next to their face as indicated by an acetate graph on the LCD feedback monitor of the video camera. Speakers started with their mouths closed and hands lowered, and returned to the rest position between words. No hand gestures were made for the AF condition. The Mandarin characters and Pinyin romanizations were presented to the speakers via PowerPoint slides. Videos were captured on a high definition camcorder (Canon Vixia HF30) at a recording rate of 30fps. Concurrent high quality audio was recorded using a Shure KSM109 microphone at 48kHz.

2.2.3. **Stimulus Editing**

The videos were edited using Final Cut Pro X to contain one word per stimulus, with the separately recorded high quality audio replacing the audio track captured by the on-camera microphone.

Using an automated FFMPEG script, the audio track for each stimulus video was normalized to 65dB SPL. Each of the audio stimuli was embedded in cafeteria noise to make the tonal information more difficult to perceive, with the goal of inducing partial visual reliance. To determine an optimal signal-to-noise ratio (SNR), Mandarin and English pilot subjects were tested on a smaller subset of audio-only stimuli embedded in +10, +5, 0, -5, -10, -12, -15 and -18dB, noise, with the goal of inducing 30% error in native Mandarin perceivers, and a 60% error rate in non-native English perceivers. At -12dB, the error rate was 15% for the Mandarin group and 54% for the English group. Most importantly, the tonal information remained audible to most perceivers at an SNR of -12dB without being completely masked by noise, which was the case at -15dB. Consequently, the 65dB SPL audio track for each stimulus was embedded in 77dB SPL cafeteria noise using FFMPEG.

The videos were mirrored horizontally so that the tone contour trace in AFG videos would travel left to right for perceivers during the experiment. AF videos were also mirrored for consistency. Each video was 4 seconds long to ensure that all the articulatory and gestural movements were captured.

For each modality, syllable and speaker, each auditory tone was paired with a tone-congruent video as well as the three other tone-incongruent videos, producing four tone-congruent pairings (one for each tone) and 12 tone-incongruent pairings (with all the possible audio and visual tone pairings differing in tone, e.g., audio-Tone1 + video-Tone2). Thus, for example, a tone-incongruent AFG auditory level Tone 1, visual rising Tone 2 (AFG-A1V2) stimulus would contain the visual track from the original AFG Tone 2 recording paired with the auditory track from the AFG Tone 1 recording. All videos presented during the experiment were cross-spliced in this manner, including the tone-congruent ones, in order to keep the treatment consistent across all stimuli. To accomplish this, a tone-congruent AFG auditory Tone 1, visual Tone 1 (AFG-A1V1) stimulus would contain the visual track from the original AFG recording paired with the auditory track from the AF recording.

Additionally, to address the potential effects of durational differences across tones (particularly for tone incongruent stimuli) in the audio-video pairings, an auditory duration-modified set of stimuli was created. For each stimulus, word onsets and offsets for the audio track of each video were manually marked, and the word durations were extracted in Praat (Boersma & Weenik, 2015). For each tone pairing, the durational difference between the original and replacement audio tones was calculated, and a stretch/compression factor was applied to the replacement tone based on the original tone duration. Then the replacement audio file was stretched or compressed accordingly. These Duration Modified audio segments were then overlaid onto the original video using Final Cut Pro X, aligning the replacement audio with the word onset of the original. Thus, these duration-modified pairings were well matched for duration in the auditory and visual input. However, considering that the stretching or compressing of the audio files may affect the (spectral and temporal) naturalness of tones, the duration-unmodified condition with natural audio was also retained. Both sets of stimuli were presented to perceivers as part of the experiment.

In total, each participant perceived 384 test stimuli (192 AF, 192 AFG) over the course of two sessions. Each of the two modalities consisted of 96 incongruent trials (12 tone pairings x 2 syllables x 2 duration modification conditions x 2 speakers) and 96 congruent trials (4 tone pairings x 2 syllables x 3 repetitions x 2 duration modification conditions x 2 speakers). Moreover, 8 additional non-noise embedded audio files (4 tones x 1 syllable "*duo*" x 2 speakers) were included as tone familiarization stimuli, and 8 additional tone-congruent audio-video files (4 tones x 1 syllable "*ge*" x 2 speakers) in each modality (AF, AFG) were prepared as task practice stimuli before the experiment.

All the AF and AFG tokens were evaluated by two native Mandarin speakers for accuracy as well as for audio and video quality. The speakers correctly identified the stimuli and rated them as satisfactory exemplars of the intended tones and gestures.

2.3. **Procedures**

The experiments were conducted in sound-attenuated perception booths at the Language and Brain Lab at Simon Fraser University. Stimuli were presented using Paradigm Stimulus Presentation software (Perception Research Systems, 2007) on 15-inch LCD monitors. Video stimuli were presented at 1024x576 resolution, and audio was

presented using AKG brand circumaural headphones. Participants were scheduled for two one-hour test sessions, separated by at least one hour's break. They were compensated with their choice of $30 cash or research participation credits for their Linguistics classes.

Prior to the test sessions, participants were introduced to the Mandarin lexical tone system with the 8 tone familiarization stimuli described above, presented in an audio-only condition. They then listened to the same stimuli again in a practice response task, using the descriptors "Level", "Rising", "Dipping", and "Falling" to identify the tones by pressing the corresponding buttons on the keyboard. All participants were required to perform above chance before continuing; none had to repeat the task.

After completion of the tone familiarization exercise, participants moved on to the test sessions. Each session contained half of the stimulus set and was further divided into two test blocks by modality of presentation: AF and AFG. Each block consisted of four practice trials, followed by 96 experimental trials. The speakers, syllables, tones, tone congruency, and duration modification factors were randomized for presentation within each block. Block presentation was counter-balanced across session and participants.

The task in each experimental trial required perceivers to watch and listen to a stimulus video of a speaker producing a target word, and then respond to the question "Which tone did you perceive?" by identifying the tone as Level, Rising, Dipping, or Falling, and pressing the correspondingly labeled button on the keyboard. Perceivers were instructed to respond as quickly as possible after the response screen appeared, and were given a maximum of 4 seconds to respond. At the end of the second session, participants completed a feedback questionnaire about the experiment that included subjective questions about which portions of the stimuli (speaker voice, face, gesture, etc.) they found helpful in completing the perception task.

Chapter 3. **Results**

3.1. **Effects of Audio-visual Congruency**

First, to evaluate tone perception as a function of the congruency of auditory and visual (facial/gestural) input, tone perception accuracy was compared in the congruent and incongruent conditions. The auditory input served as the basis for tone accuracy measurements in the incongruent conditions. Figure 2 illustrates these congruency comparisons.

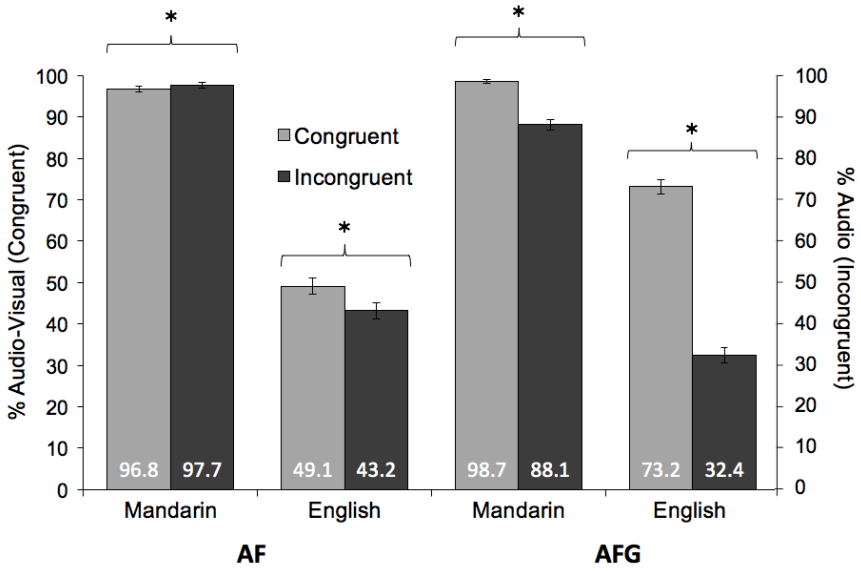

Figure 2 **Mean accuracy comparisons between Congruent and Incongruent conditions by Group (Mandarin, English) and Modality (AF, AFG)**

Congruent accuracy measured as % responses aligned with audio-visual tone; Incongruent accuracy measured as % responses aligned with auditory tone. * indicates statistically significant Congruency effects ($p < .05$). Error bars indicate 95% confidence interval.

The data were submitted to multilevel mixed effect logistic regression with Congruency (Congruent, Incongruent), Group (Mandarin, English), and Modality (AF, AFG) as fixed factors. A random effect was added on the intercept term to account for different perceivers. Factors peripheral to the focus of the study were also adjusted for in the analysis, including Duration modification (Modified, Unmodified), Tone (Level, Rising, Dipping, Falling), Speaker gender (Male, Female), Syllable (*Ye*, *You*), and Repetition (1, 2, 3). The estimated coefficients, summarized in Table 1 below, reveal significant main effects of Congruency, Group, and Modality, as well as main effects of Duration

modification and Tone. For brevity, only significant interactions involving Congruency, Group and Modality (the main factors of concern here) are reported. The significant effects involving Duration modification and Tone will be further analyzed in Sections 3.4 and 3.5, respectively.

Table 1 Summary of mixed effect logistic regression model for tone identification accuracy

Factor	Estimate	Std. Error	z-value	Wald test p-value
(Intercept)	0.92	2.21	0.42	.677
Congruency	-0.31	0.07	-4.56	< .001
Group	4.17	0.32	12.98	< .001
Modality	1.24	0.07	18.70	< .001
Duration modification	0.19	0.04	4.53	< .001
Tone	0.05	0.02	2.50	.012
Speaker gender	-0.01	0.02	-0.50	.618
Syllable	-0.01	0.04	-0.05	.958
Repetition	-0.04	0.06	-0.54	.591
Congruency x Group	0.67	0.19	3.43	.001
Congruency x Modality	-1.81	0.09	-19.25	< .001
Congruency x Group x Modality	-1.34	0.29	-4.64	< .001

As shown in Table 1, the logistic regression revealed a significant three-way interaction of Congruency x Group x Modality, as well as two-way interactions of Congruency x Group and Congruency x Modality. To further assess these interactions, likelihood ratio tests for each modality were conducted to determine whether including a Congruency x Group interaction term would improve the model fit compared to a reduced model that excluded the interaction term but retained Congruency, Group, Duration Modification, Tone, Speaker gender, Syllable, and Repetition as factors. Significant interactions of Congruency x Group were found for both the AF [$\chi^2(1) = 11.50$, $p < .001$] and AFG [$\chi^2(1) = 17.77$, $p < .001$] modalities. With the same approach, significant Congruency x Modality interactions were observed for the Mandarin [$\chi^2(1) = 155.90$, $p < .001$] and English [$\chi^2(1) = 383.64$, $p < .001$] groups.

These significant interactions motivated further comparisons of Congruency within each Modality and Group, using Wald tests. First, in the AF modality, for Mandarin perceivers, accuracy was unexpectedly higher in the incongruent condition (AF-I) than the

congruent condition (AF-C) [AF-C/AF-I = 0.61, CI = (0.42, 0.91), z = 2.51, p = .012], although it should be noted that performance was close to ceiling in both congruency conditions (Figure 2). In contrast, for the English group, tone accuracy was significantly higher in AF-C than in AF-I [AF-C/AF-I = 1.37, CI = (1.19, 1.58), z = -4.53, p < .001], showing the expected positive effects when congruent auditory and facial information were presented. The positive effects of congruency were also revealed in the AFG modality, where congruent auditory and facial-gestural input (AFG-C) produced higher tone accuracy compared to incongruent input (AFG-I) for both the Mandarin [AFG-C/AFG-I = 29.88, CI = (17.13, 52.11), z = -12.21, p < .001] and English [AFG-C/AFG-I = 8.99, CI = (7.61, 10.62), z = -26.42, p < .001] groups.

In sum, the results demonstrate more effective perception with congruent (than incongruent) auditory and visual (facial/gestural) input for both native (Mandarin) and non-native (English) perceivers, with the exception of the Mandarin AF condition where ceiling performance was observed.

3.2. **Effects of Input Modality**

To determine the extent to which the AFG modality relative to AF affected tone perception for both Mandarin and English perceivers, congruent audio-visual trials were analyzed using logistic regression with Group and Modality as fixed effects, with model adjustments for the same peripheral and random factors as reported in 3.1. A significant main effect of Modality was observed across groups, where tone identification was more accurate in AFG (86%) than in AF (73%), [AFG/AF = 3.56, CI (3.11, 4.08), z = 18.70, p < .001]. A significant main effect of Group was observed across modalities, with Mandarin perceivers (98%) outperforming English perceivers (62%), [Mandarin/English = 48.42, CI (26.05, 90.02), z = 12.38, p < .001]. No significant Modality x Group interaction was observed [$\chi^2(1) = 1.51$, p = .22].

Despite the lack of interaction, the Mandarin perceivers' near-ceiling performance in both AF and AFG conditions motivated further Wald tests comparing Modality for each Group. The results confirmed that tone accuracy was superior in AFG compared to AF for both native Mandarin perceivers [AFG/AF = 2.78, CI (1.81, 4.27), z = 4.77, p < .001] and English perceivers [AFG/AF =3.55, CI (3.10, 4.07), z = 18.67, p < .001]. A significant effect of Group for each Modality was observed, with native Mandarin perceivers outperforming

English perceivers in both AF [Mandarin/English = 47.94, CI (24.77, 92.75), z = 11.50, p < .001] and AFG [Mandarin/English = 44.25, CI (20.28, 96.54), z = 9.50, p < .001]. Figure 3 illustrates these differences in performance by Modality and Group.

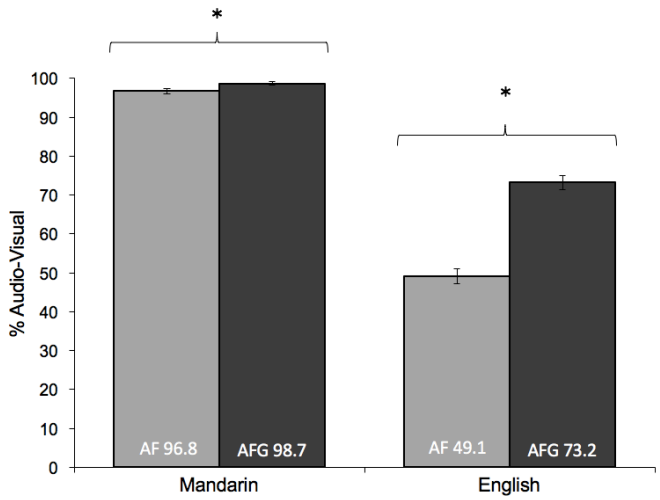

Figure 3 **Mean accuracy comparisons for audio-visual tone congruent trials by Group (Mandarin, English) and Modality (AF, AFG)**

* indicates statistically significant congruency effects (p < .05). Error bars indicate 95% confidence interval.

These results indicate the benefit of gesture, where AFG produced higher tone identification rates compared to AF for both native (Mandarin) and non-native (English) groups. The perceptual benefits of gesture were more pronounced for the non-native group, as the native perceivers achieved very high tone identification accuracy rates with and without gesture, as expected.

3.3. Effects of Perceptual Weighting of Auditory and Visual Input

To quantify the effects of visual (facial and gestural) relative to auditory information on perception in incongruent AF and AFG conditions, a perceptual weighting analysis in the present section sorted perceiver responses for each token into three categories: correct response based on auditory tone input (A), correct response based on visual tone input (V), or one of the remaining (Other, O) two tones (since participants were given all four tones as response options). For example, in the case of a token consisting of an audio Rising tone cross-spliced with a visual Falling tone, a Rising response would be coded as

A, Falling as V, and a Level or Dipping tone response as O. Figure 4 shows the varying proportion of responses by Modality and Group, categorized as A, V, and O, for the Mandarin and English perceiver groups in AF and AFG conditions.

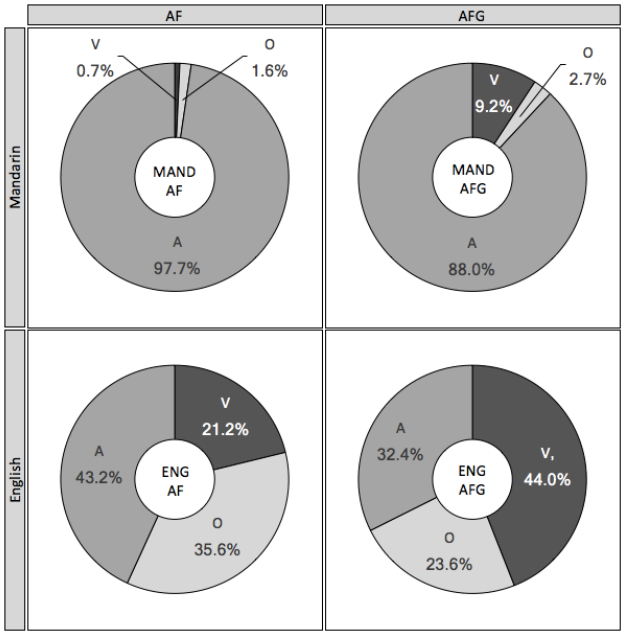

Figure 4 **Group (Mandarin, English) and Modality (AF, AFG) comparisons of Incongruent data, classified by Audio, Visual (facial/gestural), or Other response type.**

A: % correct responses based on audio tone input
V: % correct responses based on visual tone input
O: % other tone responses
MAND: Mandarin group, ENG: English group

For each Group and Modality, Friedman's tests with subsequent Wilcoxon-Nemenyi-McDonald-Thompson post-hoc tests were conducted to determine the rank order of the participant responses in the A, V, and O response categories. Within-group and modality weighting of A and V proportions were then evaluated using pairwise t-tests. These within-group proportions were subsequently submitted to two-sample t-tests to determine the differences in perceptual weighting between modalities.

For Mandarin perceivers in the AF modality, Friedman's test results indicated that the A, V, and O response categories were not equally preferred [$\chi^2(2) = 11.69$, $p < .001$]. As expected, the proportion of A-based responses was significantly greater than both V- and O-based responses (*ps*. < .001), whereas the latter two categories did not differ

significantly (p = .306). Likewise, in AFG, significant differences among response categories were also observed [$\chi^2(2) = 13.06$, $p < .001$], with responses to A significantly outweighing V, which in turn outweighed O (*ps*. < .001). However, between-modality comparisons using two sample t-tests showed that the A-based response was significantly greater in AF than in AFG [$t(25) = 2.49$, $p = .020$], whereas the V-based response was significantly greater in AFG than AF [$t(25) = 2.31$ $p = .029$].

For the English group, the AF condition also revealed significant differences in audio-visual weighting [$\chi^2(2) = 10.61$, $p < .001$], with post-hoc tests indicating greater A-based responses over O, which in turn significantly outranked V-based responses (*ps*. < .001). In the AFG condition, significant differences between category responses were observed as well [$\chi^2(2) = 11.69$, $p < .001$]. However, in contrast to the other results of A-dominant response patterns, with gesture, English perceivers' responses following the visual input increased to the extent that V exceeded both the A and O categories (*ps*. < .001); while the latter two did not differ ($p = .079$). Comparisons between the AF and AFG modalities showed that English perceivers' A response was significantly greater in AF than AFG [$t(25) = 4.63$, $p < .001$], whereas their V response was significantly greater in AFG than in AF [$t(25) = 8.67$, $p < .001$].

The analyses thus far have indicated an inverse relationship between the variables of auditory and visual response, where A decreases when V increases. Pearson correlation coefficients were calculated to determine the linearity of this relationship in each Group and Modality. Overall, strong negative correlations were found for Audio and Visual tone responses for all groups and modalities in the Incongruent conditions. In the Mandarin group, significant negative correlations were found for both AF ($r = -0.97$, $n = 26$, $p < .001$) and AFG ($r = -0.98$, $n = 26$, $p < .001$). Similarly, in the English group, there were significant negative correlations in AF ($r = -0.88$, $n = 26$, $p < .001$) as well as in AFG ($r = -0.88$, $n = 26$, $p < .001$). Figure 5 illustrates these relationships by plotting % Audio Tone against % Visual Tone for each Group, Modality, and Subject.

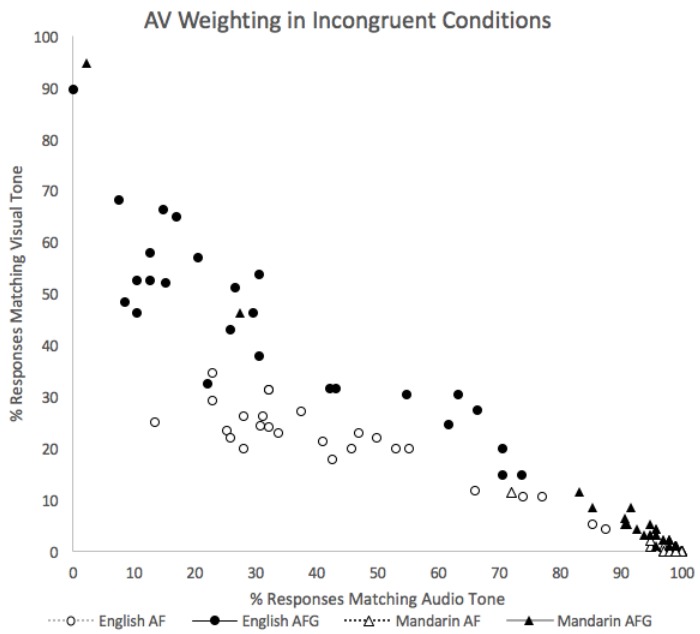

Figure 5 **Correlation between auditory and visual response for each Group and Modality pairing (English AF, English AFG, Mandarin AF, Mandarin AFG) in the Incongruent conditions**

The differences in correlation coefficients between groups motivated further analyses to test whether these increases in V weighting between the AF and AFG modality were significantly different between groups. Cross-group comparisons using pairwise t-tests showed that both Mandarin [$t(25) = 2.31$, $p = .029$] and English perceivers [$t(25) = 8.67$, $p < .001$] significantly increased their V weighting from AF to AFG. Within each group, the differences for audio and visual responses between AF and AFG were then calculated for each subject. A two sample t-test of the differences [$t(50) = 3.18$, $p = .003$] revealed that the increase in visual weighting with the inclusion of gesture in the AFG condition was greater for English perceivers (22.8%), than for Mandarin perceivers (8.5%). Figure 6 illustrates these differences below.

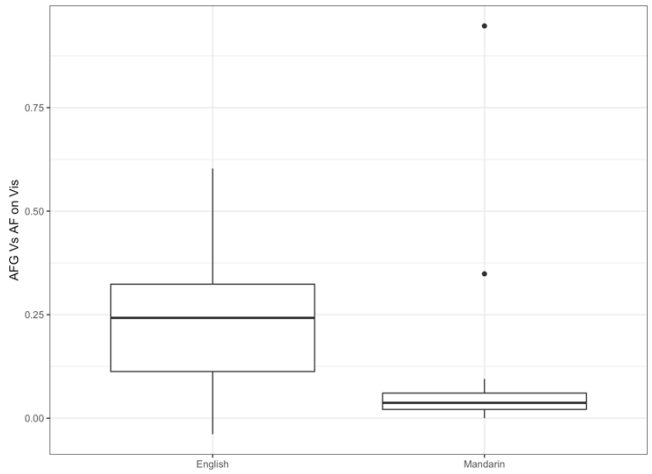

Figure 6 Increase in visual weighting from AF to AFG by Group (English, Mandarin)

To summarize, the analysis of the incongruent data showed that in stimuli where auditory and visual cues for tone were mismatched, both native (Mandarin) and non-native (English) perceivers increased their visual weighting when highly salient gesture cues were available in the AFG modality (as compared to AF). Furthermore, the non-native group weighted the visual tone even more highly than the auditory tone input when gestures were present.

3.4. Effects of Duration of the Auditory and Visual Input

As discussed in section 2.2.2, two sets of stimuli were created for the incongruent stimuli: the duration-modified set with modified audio tone duration to match the duration of the visual tone, and the duration-unmodified set with the natural audio tone duration retained. The main effect of Duration modification in the full model logistic regression in 3.1 motivated further analysis on the incongruent data to determine if durational congruency affects perception as a function of Modality and Group. A likelihood ratio test between the full model (including all two and three way interactions) and the reduced model excluding the interaction term indicated no significant Group x Modality x Duration modification interaction for either the Auditory-based responses [$\chi^2(1) = 0.60$, $p = .440$] or the Visual-based responses [$\chi^2(1) = 0.7244$, $p = .395$]. This result indicated that duration modification affected all groups and modalities in the same way, and therefore no further analysis was undertaken.

3.5. **Effects of Individual Tones**

The significant main effect of Tone (p = .012) observed in the full model logistic regression in section 3.1 motivated additional analyses of potential individual tone effects as functions of Modality and Group, for both the audio-visual congruent and incongruent data. First, likelihood ratio tests between the full model and the reduced model, which excluded the interaction term, were used to assess the Tone x Modality x Group interactions. If significant interactions were found, further Friedman's tests with Wilcoxon-Nemenyi-McDonald-Thompson post-hoc tests were then employed to tease apart the differing effects of individual tones in each modality and for each group. Figure 7 illustrates individual tone perception in AF and AFG for Mandarin and English perceivers in terms of (a) percent correct identification in the congruent conditions, (b) percentage of responses matching the auditory tone in the incongruent conditions, and (c) percentage of responses matching the visual tone in the incongruent conditions.

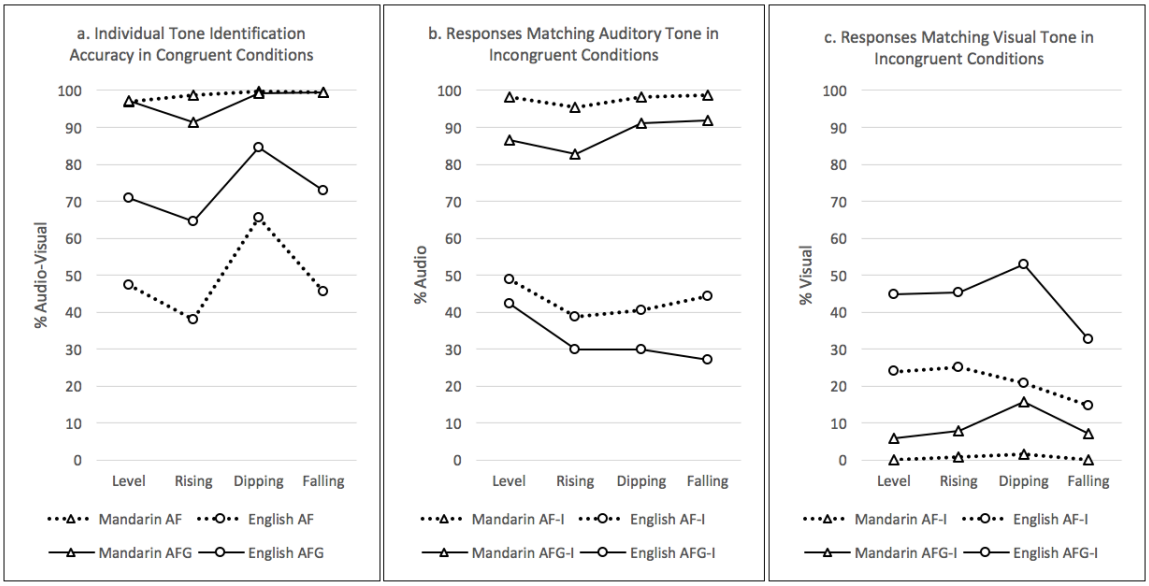

Figure 7 **Individual tone (Level, Rising, Dipping, Falling) perception by Group (Mandarin, English) and Modality (AF, AFG)**

(a) Percent correct identification in audio-visual congruent conditions.
(b) Percentage of responses matching the auditory tone in incongruent conditions.
(c) Percentage of responses matching the visual tone in incongruent conditions.

Likelihood ratio tests of the Congruent data (Figure 5a) revealed a significant interaction of Group x Modality x Tone [$\chi^2(3) = 16.561$, $p < .001$]. Although identification accuracy for each tone was generally very high in the Mandarin perceiver group, Friedman's test showed significant differences in the AF condition [$\chi^2(3) = 32.92$, $p < .001$], with post-hoc pairwise comparisons indicating slightly (although significantly) higher accuracy for Level, Dipping, and Falling tones than Rising tone (*ps.* < .001). A significant tone effect was also observed in AFG [$\chi^2(3) = 10.61$, $p = .014$], showing better perception of Dipping and Falling tones than Level tone (*ps.* = .031). A significant effect of tone was also present for the English group in the AF condition [$\chi^2(3) = 16.27$, $p = .001$], with Dipping tone performing better than both Rising and Falling tones (*ps.* $\leq$ .017). Likewise, in AFG, the significant tone effect [$\chi^2(3) = 13.49$, $p = .004$] was due to better identification of Dipping than Rising tone ($p = .002$).

Analysis of the incongruent data based on audio responses (Figure 5b) only revealed a significant Group x Tone interaction [$\chi^2(3) = 55.96$, $p < .001$]. Across modalities, significant differences between tones were found for the Mandarin group [$\chi^2(3) = 21.01$, $p < .001$], but not for the English group [$\chi^2(3) = 2.16$, $p = .130$]. For the Mandarin perceivers, Dipping tone outperformed Rising tone, and Falling tone outperformed both Level and Rising tones (*ps.* $\leq$ .024).

The video response analysis of the incongruent data (Figure 5c) also only revealed a significant Group x Tone interaction [$\chi^2(3) = 38.76$, $p < .001$]. Across modalities, significant differences between tones were found in both Mandarin [$\chi^2(3) = 32.30$, $p < .001$] and English [$\chi^2(3) = 17.63$, $p < .001$]. For the Mandarin group, Dipping tone responses were greater than all the other tones (*ps.* $\leq$ .003). For the English group, Level, Rising, and Dipping tones all outperformed Falling tone (*ps.* $\leq$ .020). Moreover, the Modality x Tone interaction was also significant [$\chi^2(3) = 25.97$, $p < .001$]. Across groups, significant differences between tones were found in both AF [$\chi^2(3) = 3.18$, $p = .008$] and AFG [$\chi^2(3) = 5.69$, $p < .001$] modalities. In AF, Rising was better than Falling tone ($p = .008$), and in AFG, Dipping was better than all the other tones (*ps.* < .001).

Overall, the most notable result of the individual tone analysis was that Dipping tone was frequently observed to outperform the other tones on the measures that included a visual component, especially in the AFG modality.

3.6. **Summary**

Taken together, the results showing better performance with congruent (than incongruent) auditory and visual (facial/gestural) input across groups indicate that perceivers can make cross-modal associations between acoustic, visual articulatory, and spatial pitch information. Importantly, the data show that these associations are not caused by durational congruency effects. Furthermore, increased perceptual accuracy and visual weighting were observed across all tones with the addition of gestural input, especially for Dipping tone. Overall, more pronounced visual benefits, particularly gestural benefits, were observed for the non-native (English) perceiver group than the native (Mandarin) group.

Chapter 4. Discussion and Conclusion

4.1. Facial Effects

The AF modality's primary purpose in this study was to facilitate direct comparisons of facial effects between groups and congruency conditions. As expected, native Mandarin perceivers outperformed tone-naïve English perceivers at tone identification. Stronger effects of facial information were revealed for perceivers who had less experience with the Mandarin language. The following discussions reflect in greater detail how our findings align with our hypotheses regarding Congruency, AV Weighting, and Tone, and how the present results compare with prior studies on similar topics.

4.1.1. Congruency

An unexpected finding of the Congruency comparisons was that Native Mandarin perceivers were less performant with congruent auditory and facial information in AF-C than in with incongruent information in AF-I. Conversely, tone identification in the English group was better in AF-C than AF-I, as hypothesized. Previous studies of audiovisual Mandarin tone perception have also found that facial cues for tone are more likely to be used by non-native perceivers who find themselves in a challenging phonetic situation, than by native perceivers (Chen and Massaro, 2008; Smith and Burnham, 2012). Although Mandarin tone contours are still audible at the SNR of -12 dB used in the current study (Mixdorff et al., 2005), it has previously been demonstrated that attention is increasingly shifted from the eyes to the mouth with increased auditory noise (Vatikiotis-Bateson et al., 1998). The differences between AF-C and AF-I conditions, significant for both groups, albeit in opposite directions, suggest that subjects were able to translate specific patterns of head and facial motion, such as head dips, jaw opening, and lip closing (Attina et al., 2010) as well as their associated durational differences (Smith and Burnham, 2012) into cross-modal cues for tone, but may have employed different strategies to do so, depending on their language experience.

Although manipulating audiovisual congruency is now an established paradigm for McGurk experiments (McGurk and Macdonald, 1976), the technique had not been applied to lexical tone in the field of audiovisual speech perception research until the present study. We hypothesized that tone identification accuracy rates would be affected by

modifying the congruency between auditory tone and facial information. Specifically, if perceivers are able to effectively incorporate facial cues for tone, then correct tone identification rates would be higher when facial cues matched audio cues than when the cues were mismatched in the incongruent condition. The results of the AF modality partially support this hypothesis. The dissimilar patterns of tone identification between groups indicate that conflicting inputs may have been processed differently by the different groups. While the Mandarin AF results do not directly support our hypothesis, their higher performance in AF-I suggests that crossmodal conflict created by mismatched cues may have counterintuitively acted to strengthen their reliable native auditory tone judgements. When the same information was presented to non-native perceivers who lack strong tone categories, facial information may have been more likely to be perceptually fused to the auditory stream. Incongruent trials may have been less likely to be perceived as intentionally misleading, especially since congruent and incongruent trials were randomized within blocks. In face-to-face conversation, there is no reasonable expectation of hearing speech sounds that are in direct opposition to their speaker's articulatory configurations, so both the native and non-native performance patterns can be reasonably accounted for in strategically recruiting complementary visual cues to support tone perception.

4.1.2. **AV Weighting**

In the AF-I condition, native Mandarin responses were 97.7% Audio and 0.7% Visual, while English perceivers' responses were 43.2% Audio, and 21.2% Visual. Consistent with previous findings (Mixdorff et al., 2005), the Mandarin group was very confident in their auditory tone intuitions, and thus, weighted the auditory information highly compared to the visual information. The English group, less adept at tone identification, as judged by the lower overall percentage accounted for by either Audio or Visual, more readily incorporated facial information, as in prior studies (Smith & Burnham, 2012), to disambiguate between tones. Within the AF modality, our results support our hypothesis that the weighting of auditory and visual information would vary depending on the language background of the perceivers. The weighting of cues in the AF modality may also be influenced by their visual salience. Previous studies such as Hazan et al. (2006) have found that the weighting of cues depended on the visual salience of the segmental contrast being perceived. The visual components attributable exclusively to tone do not

reliably involve labiality (e.g., place of articulation, lip spreading, lip rounding, lip protrusion as when articulating segmental contrasts), except to mark duration with jaw opening and lip closure (Attina et al. 2010), and so, may not have been salient enough to affect the native Mandarin group.

4.1.3. **Tone**

Categorizing the results by individual tone in AF-C, the only significant tone for the Mandarin perceivers was Rising tone, which was less accurately identified than the other three tones. For English perceivers, Dipping tone was significantly more accurate than both Rising and Falling tones. Analyzed from a visual response perspective, English perceivers in AF-I chose Rising Visual Tone (VisTone) more frequently than Falling VisTone. In relation to the visual saliency of individual tones, perceivers might grant a highly visually salient tone more visual weight than the others when encountering it during the experiment. Dipping tone is the candidate with the most visually salient features: long duration relative to other tones (Xu, 1997) with a head dipping (Smith & Burnham, 2012; Chen & Massaro, 2008) or jaw lowering motion corresponding to the articulatory configuration for the turning point of the tone. In AF-C, Dipping tone was indeed more well identified than Rising or Falling tones, indicating the possible helpfulness of the visual information; however, in AF-I, Dipping tone was not the most frequent visual response – only Rising VisTone showed a significant effect over the lowest visual response – Falling VisTone.

4.2. **Gestural Effects**

Overall, native Mandarin perceivers outperformed English perceivers in the AFG modality. However, the highly salient tone gestures provided illustrative aids that influenced the responses of both groups in both congruent and incongruent conditions. The following sections will discuss how gesture may have affected tone perception in ways that facial information alone did not.

4.2.1. **Congruency**

Previous studies of word-level iconic and beat gestures in L2 perception (Hirata & Kelly, 2010; Hirata et al., 2014; Kelly et al., 2014, 2009; Kelly & Lee, 2012) have shown

that similar types of gestures can be unhelpful when paired with a learning task, due to increased processing loads. Perceivers of Mandarin tones have also been shown to learn certain tones more slowly when training occurred with gestures than without (Eng et al., 2014; Morett & Chang, 2015). In the present study, where processing loads were lower due to the absence of a learning component, both groups of perceivers were more accurate in AFG-C than AFG-I, showing that perceivers were able to make the crossmodal connection between the tone gesture and the auditory tone. This finding compares well with a recent study of Japanese intonational contrasts (Kelly, Bailey, & Hirata, 2017) which utilized a similar experimental paradigm with the addition of a no-gesture condition, but explicitly instructed participants to only pay attention to the audio. Their study also found that for intonation, congruent gesture resulted in greater accuracy, and incongruent gesture in lower accuracy, compared to the no-gesture condition. One final point of discussion for gestural congruency is that in the AFG-I condition, the highly salient tone gestures travel at least part of the time in the opposite direction of the acoustic frequency. A study in the same vein, Connell et al., (2013), also showed that pitch perception could be swayed upwards or downwards in the direction of the gesture. It is possible that incongruent tone gestures in the present study may have had a similar effect on tone perception. It is apparent that the less certain perceivers are about what they hear, the more they prefer to rely upon gesture as supportive information, even though such information may actually be in conflict with the auditory tone. Thus, our second hypothesis of crossmodal audiospatial linking between acoustic pitch and the visuospatial representation of pitch is further supported.

4.2.2. **AV Weighting**

There are several sets of comparisons that provide evidence for increased visual weighting as an effect of adding tone gestures to the visual stimuli. First, in comparing AFG-C and AF-C modality conditions, both native and non-native perceivers were able to identify tones more accurately when gestural input was available. Then, comparing AFG-C with AFG-I, both groups performed more accurately when the audio and gestural inputs consisted of matching tonal information, compared to when the gesture drew a different contour than the auditory tone, which suggests that visual information was highly weighted when available. A comparison of visual responses between AFG-I and AF-I revealed that the higher proportion of visual responses in AFG-I was statistically significant in both the

Mandarin and English groups. For the English group, the proportion of visual responses in AFG-I was also significantly larger than Audio responses. Taken together, these results provide a fairly clear picture that tone gestures are highly salient visual contrasts, and greatly affect visual weighting (Hazan et al., 2006). In the present study, such high visual weighting effectively decreased overall performance in AFG for both groups compared to AF. However, when considered with the perspective that manual gestures are generally used to provide redundant cues to concurrent speech (Hostetter, 2011), and that a collocutor is highly unlikely to make metaphoric gestures that are intentionally misleading (unlike the experiment), the strategy of weighting such gestures highly is not unreasonable, and can be very effective, as in English AFG-C. Evidently, perceivers are crossmodally relating visuospatial tone gesture information to the auditory tone information.

4.2.3. **Tone**

With the addition of gesture, different patterns emerged for individual tone identification. For Mandarin AFG-C, Dipping and Falling tones were more accurately perceived than Level tone. In Mandarin AFG-I, Falling retained its accuracy over Level tone, but also became more accurate than Rising tone. Dipping tone was no longer more accurate than level tone, but was instead more accurate than Rising tone. Response patterns changed for non-native perceivers as well: in AFG-C, Dipping tone was more accurately perceived than Rising tone, while in AFG-I, the only difference was higher performance for Level tone over Falling tone. In a previous study of tone gesture, Eng et al. (2014) found a slower learning curve for Falling tone in the Audio-Face-Gesture condition compared to Audio-Only, Audio-Face, and Audio-Gesture, and a slower learning curve for Rising and Dipping tones in the Audio-Gesture condition compared to the Audio-Face condition, possibly due to the similarity of the gestural trajectories of Rising and Dipping tones. In the present study, we find similar patterns of confusion occurring in AFG-C in the disparity between Rising and Dipping tone accuracy rates, as well as in Falling tone being the least reliably identified in the presence of distracting incongruent gesture. From a visual response perspective, Mandarin perceivers in the AFG-I condition selected Dipping VisTone more frequently than Level, Rising, or Falling VisTones, while English perceivers in AFG-I selected Dipping VisTone more frequently than Falling VisTone. The duration and contour of the Dipping tone gesture correlate well with the assertion that high

salience visual contrasts have more potential to affect visual weighting (Hazan et al., 2006). Overall, the AFG results of the present study demonstrate that highly salient tone gestures captured the attention of both native and non-native speakers and impacted the accuracy of their tone judgements significantly more than facial movements alone.

4.3. General Discussion and Concluding Remarks

Recall our hypothesis that if perceivers are making crossmodal, non-arbitrary, audiospatial correspondences between acoustic and visual tonal cues and binding them during perception, then the manipulation of congruency should reveal how these links occur. For non-native perceivers who would be more likely to seek complementary or redundant input through facial or gestural sources, congruent AV information was observed to aid tone identification, while incongruent AV cues conversely inhibited tone perception. In addition, congruency was more influential to tone perception in the AFG modality than in AF, demonstrating that audiospatial correspondences were made between pitch and gesture. For the English group, within the incongruent condition, the dominant sensory choice changed between the AF and AFG modalities to favour the Visual information in AFG. Furthermore, for these non-native perceivers, the total percentage of responses accounted for by audio and visual tone increased when gesture was included in the stimuli, showing that perceivers actively recruited visual information for tone judgements.

Some unexamined yet pertinent issues peripheral to our primary research questions can now be briefly addressed here. Regarding the visual information available to perceivers, how did we know participants were looking at the gesture or the speaker's face? What if participants decided to completely ignore the visual or auditory stream in the stimuli? Questions including these were quite seriously considered during the initial design stages of the study. We considered the use of an eye-tracker, but the type of data we would have gained over a pure perception experiment did not justify the additional investment required, without a major experimental redesign. To diminish the concern that perceivers might stop looking at the visual stimuli, the verbal and written instructions in our experiment emphasized that perceivers should pay attention to *both* the auditory *and* visual information. We chose to direct their attention in this manner because as a multimodal experiment with auditory noise, foreign speech sounds, and conflicting inputs, the temptation can be strong for participants to simplify the task by shutting their eyes or

looking away from the screen, thus undermining the experiment's design. However, a recent similar study of gesture and intonation contrasts (Kelly, Bailey, & Hirata, 2017) instead instructed their participants to only pay attention to the auditory information present in the stimuli. They observed effects of gesture on auditory perception despite these instructions, so future studies may choose to adjust their instructions accordingly, knowing the outcome of both types of instructions on perception.

The analysis of the data generated by the present study only skims the surface of inquiry into specific patterns of tone fusion and confusion. There is need to develop statistically reliable methods for quantifying such nebulous percepts as "tone fusion" and "tone confusion" with distance functions, as these conflicts may be more dynamic than segmental mismatches of place of articulation or voicing. An interesting future direction would be to study the cases that were categorized as "Other" in the Incongruent conditions, as these responses did not align with either the auditory or facial/gestural tone information. Given a sufficiently large dataset, there is reason to believe that it is within this *Other* category where the most interesting perceptual fusion and conflict patterns will be discovered.

Chapter 5. **References**

Alsius, A., Navarra, J., Campbell, R., & Soto-Faraco, S. (2005). Audiovisual integration of speech falters under high attention demands. *Current Biology*, *15*(9), 839-843. doi:10.1016/j.cub.2005.03.046

Apfelstadt, H. (1988). What makes children sing well? *Applications of Research in Music Education*, *7*(1), 27-32. doi:10.1177/875512338800700108

Attina, V., Gibert, G., Vatikiotis-Bateson, E., & Burnham, D. (2010). Production of Mandarin lexical tones: Auditory and visual components. *Proceedings of AVSP-2010*, paper S4-2. Retrieved from http://www.isca-speech.org/archive/avsp10/index.html

Barnett, M. A. (1983). Replacing teacher talk with gestures: Nonverbal communication in the foreign language classroom. *Foreign Language Annals*, *16*(3), 173-176. doi:10.1111/j.1944-9720.1983.tb01446.x

Barsalou, L. (2008). Grounded cognition. *Annual Review of Psychology*, *59*, 617-645. doi:10.1146/annurev.psych.59.103006.093639

Bernstein, I. H., & Edelstein, B. A. (1971). Effects of some variations in auditory input upon visual choice reaction time. *Journal of Experimental Psychology*, *87*(2), 241-247. doi:10.1037/h0030524

Biau, E., & Soto-Faraco, S. (2013). Beat gestures modulate auditory integration in speech perception. *Brain and Language*, *124*(2), 143-152. doi:10.1016/j.bandl.2012.10.008

Boersma, P. & Weenink, D. (2015). Praat: doing phonetics by computer (Version 5.4) [Computer software]. Retrieved from http://www.praat.org/

Burnham, D., Ciocca, V., & Stokes, S. (2001). Auditory-visual perception of lexical tone. *Proceedings of EUROSPEECH-2001*, 395-398. Retrieved from http://www.isca-speech.org/archive/eurospeech_2001/

Burnham, D., Lau, S., Tam, H., & Schoknecht, C. (2001). Visual discrimination of Cantonese tone by tonal but non-Cantonese speakers, and by non-tonal language speakers. *Proceedings of AVSP-2001*, 155-160. Retrieved from http://isca-speech.org/archive_open/avsp01/

Burnham, D., Reynolds, J., Vatikiotis-Bateson, E., Yehia, H., & Ciocca, V. (2006). The perception and production of phones and tones: The role of rigid and non-rigid face and head motion. *Proceedings of ISSP 2006*, paper 14. Retrieved from http://www.cefala.org/issp2006/cdrom/main_index.html

Casasanto, D., Phillips, W., & Boroditsky, L. (2003). Do We Think About Music in Terms of Space? Metaphoric Representation of Musical Pitch. *Proceedings of CogSci 2003*, 1323. Retrieved from http://csjarchive.cogsci.rpi.edu/proceedings/2003/pdfs/255.pdf

Cavé, C., Guaitella, I., Bertrand, R., Santi, S., Harlay, F., & Espesser, R. (1995). About the relationship between eyebrow movements and Fo variations. *Proceedings of ICSLP 96*, 2175-2178. doi:10.1109/ICSLP.1996.607235

Chao, Y. R. (1968). *A grammar of spoken Chinese*. Berkeley, CA: University of California Press.

Chen, G. (1974). The pitch range of English and Chinese speakers. *Journal of Chinese Linguistics*, *2*(2), 159-171. Retrieved from http://www.cuhk.edu.hk/journal/jcl/

Chen, T. H., & Massaro, D. W. (2008). Seeing pitch: visual information for lexical tones of Mandarin-Chinese. *The Journal of the Acoustical Society of America*, *123*(4), 2356-66. doi:10.1121/1.2839004

Chen, Y., & Hazan, V. (2007). Language effects on the degree of visual influence in audiovisual speech perception. *Proceedings of ICPhS XVI*, 2177-2180. Retrieved from http://www.icphs2007.de/

Connell, L., Cai, Z. G., & Holler, J. (2013). Do you see what I'm singing? Visuospatial movement biases pitch perception. *Brain and Cognition*, *81*(1), 124-130. doi:10.1016/j.bandc.2012.09.005

Davis, C., & Kim, J. (2004). Audio-visual interactions with intact clearly audible speech. *The Quarterly Journal of Experimental Psychology Section A*, *57*(6), 1103-1121. doi:10.1080/02724980343000701

Ernst, M. O. (2007). Learning to integrate arbitrary signals from vision and touch. *Journal of Vision*, *77*(5), 1-14. doi:10.1167/7.5.7

Ernst, M. O., & Bülthoff, H. H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Sciences*, *8*(4), 162-169. doi:10.1016/j.tics.2004.02.002

Fujisaki, W., & Nishida, S. (2007). Feature-based processing of audio-visual synchrony perception revealed by random pulse trains. *Vision Research*, *47*(8), 1075-1093. doi:10.1016/j.visres.2007.01.021

Gluhareva, D., & Prieto, P. (2016). Training with rhythmic beat gestures benefits L2 pronunciation in discourse-demanding situations. *Language Teaching Research*. Advance online publication. doi:10.1177/1362168816651463

Gullberg, M. (2006). Some reasons for studying gesture and second language acquisition (Hommage à Adam Kendon). *IRAL - International Review of Applied Linguistics in Language Teaching*, *44*(2), 103-124. doi:10.1515/IRAL.2006.004

Hazan, V., Sennema, A., Faulkner, A., Ortega-Llebaria, M., Iba, M., & Chung, H. (2006). The use of visual cues in the perception of non-native consonant contrasts. *The Journal of the Acoustical Society of America*, *119*(3), 1740-1751. doi:10.1121/1.2166611

Hazan, V., Kim, J., & Chen, Y. (2010). Audiovisual perception in adverse conditions: Language, speaker and listener effects. *Speech Communication*, *52*(11-12), 996-1009. doi:10.1016/j.specom.2010.05.003

Hirata, Y., & Kelly, S. D. (2010). Effects of lips and hands on auditory learning of second-language speech sounds. *Journal of Speech, Language, and Hearing Research*, *53*(2), 298-310. doi:10.1044/1092-4388(2009/08-0243)

Hirata, Y., Kelly, S. D., Huang, J., & Manansala, M. (2014). Effects of hand gestures on auditory learning of second-language vowel length contrasts. *Journal of Speech, Language, and Hearing Research*, *57*(6), 2090-2101. doi:10.1044/2014_JSLHR-S-14-0049

Hostetter, A. B. (2011). When do gestures communicate? A meta-analysis. *Psychological Bulletin*, *137*(2), 297-315. doi:10.1037/a0022128

Houlahan, M., & Tacka, P. (2008). *Kodály today: A cognitive approach to elementary music education*. New York, NY: Oxford University Press.

Jongman, A., Wang, Y., & Kim, B. H. (2003). Contributions of Semantic and Facial Information to Perception of Nonsibilant Fricatives. *Journal of Speech, Language, and Hearing Research*, *46*(6), 1367-1377. doi:10.1044/1092-4388(2003/106)

Kelly, S., Bailey, A., & Hirata, Y. (2017). Metaphoric gestures facilitate perception of intonation more than length in auditory judgments of non-native phonemic contrasts. *Collabra: Psychology 3*(1), 7. doi:10.1525/collabra.76

Kelly, S. D., & Lee, A. L. (2012). When actions speak too much louder than words: Hand gestures disrupt word learning when phonetic demands are high. *Language and Cognitive Processes, 27*(6), 793-807. doi:10.1080/01690965.2011.581125

Kelly, S. D., McDevitt, T., & Esch, M. (2009). Brief training with co-speech gesture lends a hand to word learning in a foreign language. *Language and Cognitive Processes*, *24*(2), 313-334. doi:10.1080/01690960802365567

Kelly, S. D., Ozyürek, A., & Maris, E. (2010). Two sides of the same coin: Speech and gesture mutually interact to enhance comprehension. *Psychological Science*, *21*(2), 260-267. doi:10.1177/0956797609357327

Kelly, S., Healey, M., Özyürek, A., & Holler, J. (2015). The processing of speech, gesture, and action during language comprehension. *Psychonomic Bulletin & Review*, *22*(2), 517-523. doi:10.3758/s13423-014-0681-7

Kelly, S. D., Hirata, Y., Manansala, M., & Huang, J. (2014). Exploring the role of hand gestures in learning novel phoneme contrasts and vocabulary in a second language. *Frontiers in Psychology*, *5*(673), 1-11. doi:10.3389/fpsyg.2014.00673

Kelly, S. D., Kravitz, C., & Hopkins, M. (2004). Neural correlates of bimodal speech and gesture comprehension. *Brain and Language*, *89*(1), 253-260. doi:10.1016/S0093-934X(03)00335-3

Kelly, S. D., Ward, S., Creigh, P., & Bartolotti, J. (2007). An intentional stance modulates the integration of gesture and speech during comprehension. *Brain and Language*, *101*(3), 222-233. doi:10.1016/j.bandl.2006.07.008

Kim, J., Cvejic, E. & Davis. C. 2014. Tracking eyebrows and head gestures associated with spoken prosody. *Speech Communication*, *57*, 317-330. doi:10.1016/j.specom.2013.06.003

Krahmer, E., & Swerts, M. (2007). The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language*, *57*(3), 396-414. doi:10.1016/j.jml.2007.06.005

Kuhl, P. K., & Meltzoff, A. N. (1982). The bimodal perception of speech in infancy. *Science*, *218*(4577), 1138-1141. doi:10.1126/science.7146899

Kuhl, P. K., & Meltzoff, A. N. (1984). The intermodal representation of speech in infants. *Infant Behavior and Development*, *7*(3), 361-381. doi:10.1016/S0163-6383(84)80050-8

Küssner, M. B., Tidhar, D., Prior, H. M., & Leech-Wilkinson, D. (2014). Musicians are more consistent: Gestural cross-modal mappings of pitch, loudness and tempo in real-time. *Frontiers in Psychology*, *5*(789), 1-15. doi:10.3389/fpsyg.2014.00789

Lansing, C. R., & McConkie, G. W. (1999). Attention to facial regions in segmental and prosodic visual speech perception tasks. *Journal of Speech, Language, and Hearing Research*, *42*(3), 526-539. doi:10.1044/jslhr.4203.526

Lavie, N. (1995). Perceptual load as a necessary condition for selective attention. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(3), 451-468. doi:10.1037/0096-1523.21.3.451

Lazaraton, A. (2004). Gesture and speech in the vocabulary explanations of one ESL teacher: A microanalytic inquiry. *Language Learning*, *54*(1), 79-117. doi:10.1111/j.1467-9922.2004.00249.x

Lewkowicz, D. J., & Hansen-Tift, A. M. (2012). Infants deploy selective attention to the mouth of a talking face when learning speech. *Proceedings of the National Academy of Sciences, USA*, *109*(5), 1431-1436. doi:10.1073/pnas.1114783109

Liao, M.-Y. (2008). The effects of gesture use on young children's pitch accuracy for singing tonal patterns. *International Journal of Music Education*, *26*(3), 197-2113. doi:10.1177/0255761408092525

Liao, M.-Y., & Davidson, J. W. (2007). The use of gesture techniques in children's singing. *International Journal of Music Education*, *25*(1), 82-94. doi:10.1177/0255761407074894

Lin, W. C. J. (1985). Teaching Mandarin tones to adult English speakers: Analysis of difficulties with suggested remedies. *RELC Journal*, *16*(2), 31-47. doi:10.1177/003368828501600207

Liu, Y., Wang, M., Perfetti, C. A., Brubaker, B., Wu, S., & MacWhinney, B. (2011). Learning a tonal language by attending to the tone: An in vivo experiment. *Language Learning*, *61*(4), 1119-1141. doi:10.1111/j.1467-9922.2011.00673.x

Macleod, A., and Summerfield, Q. (1990). A procedure for measuring auditory and audiovisual speech-reception thresholds for sentences in noise: Rationale, evaluation, and recommendations for use. *British Journal of Audiology*, *24*(1), 29-43.

Marks, L. E. (1987). On cross-modal similarity: Auditory-visual interactions in speeded discrimination. *Journal of Experimental Psychology: Human Perception and Performance*, *13*(3), 384-394. doi:10.1037/0096-1523.13.3.384

McCafferty, S.G. (2006). Gesture and the materialization of second language prosody. *IRAL – International Review of Applied Linguistics in Language Teaching*, *44*(2), 197-209. doi:10.1515/IRAL.2006.008

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746-748. doi:10.1038/264746a0

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought.* Chicago, IL: University of Chicago Press.

McNeill, D. (2005). *Gesture and thought*. Chicago, IL: University of Chicago Press. doi:10.7208/chicago/9780226514642.001.0001

Mixdorff, H., Hu, Y., & Burnham, D. (2005). Visual cues in Mandarin tone perception. *Proceedings of INTERSPEECH-2005*, 405-408. Retrieved from http://www.isca-speech.org/archive/interspeech_2005/

Mixdorff, H., Wang, Y., & Hu, Y. (2008). Robustness of tonal and segmental information in noise - Auditory and visual contributions. *Proceedings of Speech Prosody 2008*, 261-264. Retrieved from http://isle.illinois.edu/sprosig/sp2008/

Morett, L. M., & Chang, L.-Y. (2015). Emphasising sound and meaning: pitch gestures enhance Mandarin lexical tone acquisition. *Language, Cognition and Neuroscience*, *30*(3), 347-353. doi:10.1080/23273798.2014.923105

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., and Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility: Head movement improves auditory speech perception. *Psychological Science*, *15*(2), 133-137. doi:10.1111/j.0963-7214.2004.01502010.x

Nielsen, K. (2004). Segmental differences in the visual contribution to speech intelligibility. *Proceedings of INTERSPEECH-2004*, 2533-2536. Retrieved from http://www.isca-speech.org/archive/interspeech_2004/

Parise, C. V., & Spence, C. (2009). “When birds of a feather flock together”: Synesthetic correspondences modulate audiovisual integration in non-synesthetes. *PLoS ONE*, *4*(5). doi:10.1371/journal.pone.0005664

Patterson, M. L., & Werker, J. F. (2003). Two-month-old infants match phonetic information in lips and voice. *Developmental Science*, *6*(2), 191-196. doi:10.1111/1467-7687.00271

Sekiyama, K., Tohkura, Y. (1993). Inter-language differences in the influence of visual cues in speech perception. *Journal of Phonetics*, *21*(4), 427-444.

Smith, D., & Burnham, D. (2012). Faciliation of Mandarin tone perception by visual speech in clear and degraded audio: Implications for cochlear implants. *The Journal of the Acoustical Society of America*, *131*(2), 1480-1489. doi:10.1121/1.3672703

Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics*, *73*(4), 971-995. doi:10.3758/s13414-010-0073-7

Sueyoshi, A., & Hardison, D. M. (2005). The role of gestures and facial cues in second language listening comprehension. *Language Learning*, *55*(4), 661-699. doi:10.1111/j.0023-8333.2005.00320.x

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America*, *26*(2), 212-215. doi:10.1121/1.1907309

Summerfield, Q. (1983). Audio-visual speech perception, lip reading and artificial stimulation. In M. E. Lutman & M. P. Haggard (Eds.), *Hearing Science and Hearing Disorders* (pp. 131-182). London: Academic Press Inc. (London) Ltd.

Swerts, M., & Krahmer, E. (2008). Facial expression and prosodic prominence: Effects of modality and facial area. *Journal of Phonetics*, *36*(2), 219-238. doi:10.1016/j.wocn.2007.05.001

Tang, L. Y. W., Hannah, B., Jongman, A., Sereno, J., Wang, Y., & Hamarneh, G. (2015). Examining visible articulatory features in clear and plain speech. *Speech Communication*, *75*, 1-13. doi:10.1016/j.specom.2015.09.008

Tsai, R. (2011). Teaching and learning the tones of Mandarin Chinese. *Scottish Languages Review*, *24*, 43-50. Retrieved from http://www.scilt.org.uk/Library/ScottishLanguagesReview/tabid/2089/Default.aspx

Vatikiotis-Bateson, E., Eigsti, I. M., Yano, S., & Munhall, K. G. (1998). Eye movement of perceivers during audiovisual speech perception. *Perception & Psychophysics*, *60*(6), 926-40. doi:10.3758/BF03211929

Wang, Y., Spence, M., & Jongman, A. (1999). Training American listeners to perceive Mandarin tones. *The Journal of the Acoustical Society of America*, *106*(6), 3649-3658. doi:10.1121/1.428217

Wang, Y., Behne, D. M., & Jiang, H. (2008). Linguistic experience and audio-visual perception of non-native fricatives. *The Journal of the Acoustical Society of America*, *124*(3), 1716-1726. doi:10.1121/1.2956483

Wang, Y., Behne, D. M., & Jiang, H. (2009). Influence of native language phonetic system on audio-visual speech perception. *Journal of Phonetics*, *37*(3), 344-356. https://doi.org/10.1016/j.wocn.2009.04.002

Welch, G. F. (1985). A schema theory of how children learn to sing in tune. *Psychology of Music*, *13*(1), 3-18. doi:10.1177/0305735685131001

Wu, Y. C. & Coulson, S. (2007). How iconic gestures enhance communication: An ERP study. *Brain and Language*, *101*(3), 234-245. doi:10.1016/j.bandl.2006.12.003

Xu, Y. (1997). Contextual tonal variations in Mandarin. *Journal of Phonetics*, *25*(1), 61-83. doi:10.1006/jpho.1996.0034